RAP: PHASED RETRIEVAL-AUGMENTED REASONING FOR INCIDENT-AWARE TRAFFIC PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

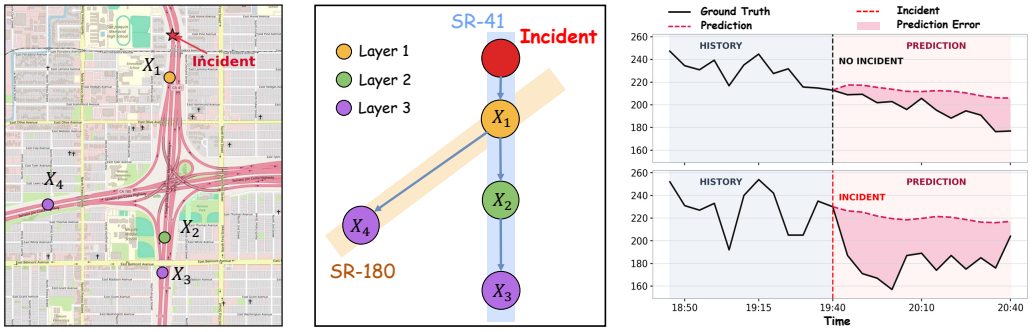
Traffic prediction plays a pivotal role in route planning and traffic management. Most existing methods focus primarily on historical spatio-temporal dependencies, overlooking external disruptions, particularly non-recurrent traffic incidents that can substantially alter traffic patterns. However, the unpredictability and heterogeneous impacts of incidents make it difficult to extract transferable incident-related knowledge from historical data, limiting the ability to accurately capture incident-affected traffic dynamics. To address these challenges, we propose RAP, a phased retrieval-augmented reasoning framework for incident-aware traffic prediction. RAP explicitly models incident-induced dynamics by integrating retrieval-augmented mechanisms with Large Language Model (LLM)-based reasoning. Specifically, it comprises two core components: Historical Knowledge Discovery and Context-Aware Phased Reasoning. The former extracts three complementary forms of transferable evidence from historical traffic data: incident patterns, normal counterfactual sequences, and experience trajectories. The latter uses an LLM to assess and consolidate heterogeneous evidence into a compact context, then partitions the prediction horizon into distinct impact phases for phase-consistent prediction. Extensive experiments on three real-world datasets demonstrate that RAP consistently outperforms state-of-the-art baselines, showing up to 5.27% performance improvement in incident-affected scenarios. The code is available at: `https://anonymous.4open.science/r/RAP-608C`.

1 INTRODUCTION

Traffic prediction is a fundamental capability for intelligent transportation systems (ITS), supporting route planning, traffic monitoring, and remote traffic management Medina-Salgado et al. (2022); Fan et al. (2026); Almousa & Lee (2025); Derrow-Pinion et al. (2021). By predicting the future conditions of road networks from observed traffic data, prediction models provide essential information for travel decision-making and traffic control Cirstea et al. (2022); Lablack & Shen (2023). However, traffic dynamics are influenced not only by recurring spatio-temporal patterns but also by external incidents, such as traffic accidents and adverse weather Gou et al. (2026). Their effects can propagate across connected roads and alter otherwise regular conditions Fan et al. (2026). Developing prediction methods that can account for these incident-induced changes is therefore important for reliable traffic management in the real-world.

Recent advances in deep spatio-temporal learning have strengthened traffic prediction by jointly modeling temporal dynamics and spatial interactions among road segments Li et al. (2017); Wu et al. (2019); Li et al. (2018); Chen et al. (2026). In particular, graph-based Yu et al. (2018); Guo et al. (2019) and attention-based models Fang et al. (2025); Yu et al. (2018) can learn correlations across connected locations while capturing recurrent patterns in traffic observations over time. By leveraging these historical regularities, existing approaches have achieved effective predictions under normal traffic conditions and established a strong foundation for data-driven traffic management Ali et al. (2025); Fang et al. (2025). Nevertheless, their performance is commonly relies on the assumption that future traffic evolves in patterns that are sufficiently represented by previously observed sequences Kougioumtzidis et al. (2025); Wang et al. (2024); Sun et al. (2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Map. (b) Directed propagation graph. (c) Traffic flow of X_4.

Figure 1: Example of foundation model prediction failure under incident conditions. (a) Incident localization and affected sensor nodes; (b) Directed propagation graph with incident-influence layers; (c) Traffic flow comparison between no-incident (top) and incident-affected (bottom) windows.

However, incident-affected traffic does not conform to the recurrent patterns captured by conventional historical modeling Gokasar et al. (2022); Fan et al. (2026). Incidents occur irregularly, can produce effects that propagate across the road network, and may have limited or uneven representation in historical training data Alzamzami et al. (2025); Eesaar et al. (2025). Consequently, models that aggregate fixed historical windows may dilute incident-related signals or treat them as noise, making it difficult to identify transferable patterns and adapt predictions to the current context Gou et al. (2026); Fan et al. (2026). As illustrated in Figure 1, an unforeseen incident first occurs near the sensor X_1 and subsequently propagates to downstream sensors across successive impact layers, including X_2, X_3, and X_4, as shown in Figures 1(a) and 1(b). For X_4, the foundation model produces a prediction primarily from the historical traffic patterns encoded during training. Because these patterns do not explicitly represent the current incident, the model produces a significant prediction error. Among such external disturbances, non-recurrent incidents are particularly challenging because their effects are difficult to infer from recurring historical patterns. Therefore, this paper focuses on traffic prediction under non-recurrent incidents.

Recent advances in large language models (LLMs) offer a promising direction for addressing this limitation Mohammadi et al. (2025); Jin et al. (2024); Zeng et al. (2023); Yehudai et al. (2026). Their semantic understanding and reasoning capabilities can help interpret heterogeneous incident descriptions and relate them to observed traffic conditions Pan et al. (2026); Zhang et al. (2025). However, the parametric knowledge of LLMs alone does not capture the network-specific traffic responses associated with individual incidents Tao et al. (2026); Zhao et al. (2025). Retrieval augmentation can provide the necessary grounding by retrieving relevant historical incidents as contextual evidence to support the analysis of how a current incident may impact regular spatio-temporal patterns Ma et al. (2026); Liang et al. (2026); Yehudai et al. (2026); Lee et al. (2025); Abootorabi et al. (2025). Thus, in this study, we focus on developing a retrieval-augmented framework for incident-aware traffic prediction. However, developing such a framework is non-trivial due to two main challenges.

Challenge I: Extracting Transferable Knowledge. Extracting transferable incident-related knowledge from historical data is challenging. Incident records typically contain only sparse event metadata Fan et al. (2026), while the traffic impacts of these incidents can vary substantially across road segments, travel directions, temporal patterns, and pre-incident traffic conditions Gou et al. (2026); Ye et al. (2023); Tran et al. (2023). Consequently, incidents with similar semantic descriptions may induce different traffic responses, while the observed traffic patterns often combine incident-induced disruptions with underlying normal traffic dynamics. This heterogeneity and entanglement of normal and incident-induced patterns make it difficult to extract incident-specific knowledge that is both relevant and transferable to the current prediction context.

Challenge II: Reliable Reasoning. Reliable reasoning about incident-affected traffic is challenging because prediction must jointly consider observed traffic data, foundation model prediction, incident metadata, and retrieved evidence. These heterogeneous sources differ in information type, granularity, and relevance, and may imply conflicting traffic adjustments Fan et al. (2026). Direct LLM reasoning may overemphasize individual evidence or mechanically transfer historical responses Tao et al. (2026); Pan et al. (2026). The central challenge is therefore to reconcile heterogeneous and potentially conflicting evidence and translate it into quantitatively grounded predictions.

To address these challenges, we introduce RAP, a phased retrieval-augmented reasoning framework for incident-aware traffic prediction that combines a spatio-temporal foundation model with LLM-based contextual reasoning. RAP consists of two key components. Historical Knowledge Discovery extracts three complementary forms of heterogeneous evidence from historical traffic data: incident patterns, normal counterfactual sequences, and experience trajectories. Specifically, RAP retrieves similar historical incident cases to identify previously observed incident-induced effects and selects incident-free traffic windows from the same target node to construct normal counterfactual sequences that characterize ordinary traffic dynamics. It further summarizes completed prediction episodes, including their contexts, predictions, observed outcomes, and reflections, into reusable experience trajectories. Together, these sources provide complementary empirical evidence for reasoning about incident-induced traffic dynamics (solving *Challenge I*). Context-Aware Phased Reasoning organizes the heterogeneous evidence into context analysis and phased prediction driven by an LLM. It first evaluates the relevance and reliability of the retrieved evidence, consolidating complementary and conflicting signals into a compact context. It then partitions the prediction horizon into distinct impact phases (e.g., disruption and recovery) and performs phase-consistent reasoning to generate reliable predictions (solving *Challenge II*).

The main contributions are summarized as follows:

- We propose RAP, a phased retrieval-augmented reasoning framework for incident-aware traffic prediction that combines a foundation model with LLM-based reasoning.
- We develop a Historical Knowledge Discovery module to extract complementary forms of heterogeneous evidence from historical traffic data, providing relevant and transferable knowledge for the current prediction context.
- We design a Context-Aware Phased Reasoning module to organize heterogeneous evidence into context analysis and phased prediction, enabling evidence-grounded and phase-consistent traffic prediction.
- Comprehensive experiments on three real-world traffic datasets demonstrate that RAP achieves state-of-the-art performance.

2 RELATED WORK

2.1 SPATIO-TEMPORAL TRAFFIC PREDICTION

Spatio-temporal traffic prediction has been a long focus of both research and industry Chen et al. (2026); Fang et al. (2025). To incorporate spatial information, GWNET Wu et al. (2019), STGODE Fang et al. (2021),DGCRN Li et al. (2023), STGCN Yu et al. (2018), BigST Han et al. (2024) and DSTAGNN Lan et al. (2022) introduced data-driven fixed adjacency matrices within graph neural networks to capture spatial dependencies. To model dynamic point-to-point spatial correlations, models such as ASTGCN Guo et al. (2019) and PatchSTG Fang et al. (2025) apply attention mechanisms to model dynamic spatial correlations. While the deep learning models have achieved remarkable success in learning from historical traffic data, their core design is oriented towards periodic traffic patterns thereby limiting their ability to model the atypical dynamics induced by abrupt external incidents Zhang et al. (2021; 2024); Fan et al. (2026).

2.2 INCIDENT-AWARE TRAFFIC PREDICTION

Traffic incidents introduce irregular disruptions that can alter local traffic conditions and propagate across connected road segments Ahmed et al. (2024); Fan et al. (2026); Gou et al. (2026); Ye et al. (2023); Tran et al. (2023). Recent resources like TraffiDent Gou et al. (2026) provide fine-grained, spatiotemporally aligned traffic-incident data, enabling detailed study of these interactions. Building on such data, incident-aware prediction methods incorporate event attributes, road-network relationships, and contextual information into spatio-temporal prediction models Fan et al. (2026); Zheng et al. (2025). However, directly encoding incident information as auxiliary features often fails to capture the heterogeneous responses associated with different locations, directions, and traffic states Tran et al. (2023); Gou et al. (2026). RAP instead retrieves relation-matched historical incidents and extracts their observed deviations from normal traffic conditions. These incident-response patterns provide contextual evidence for LLM-based reasoning.

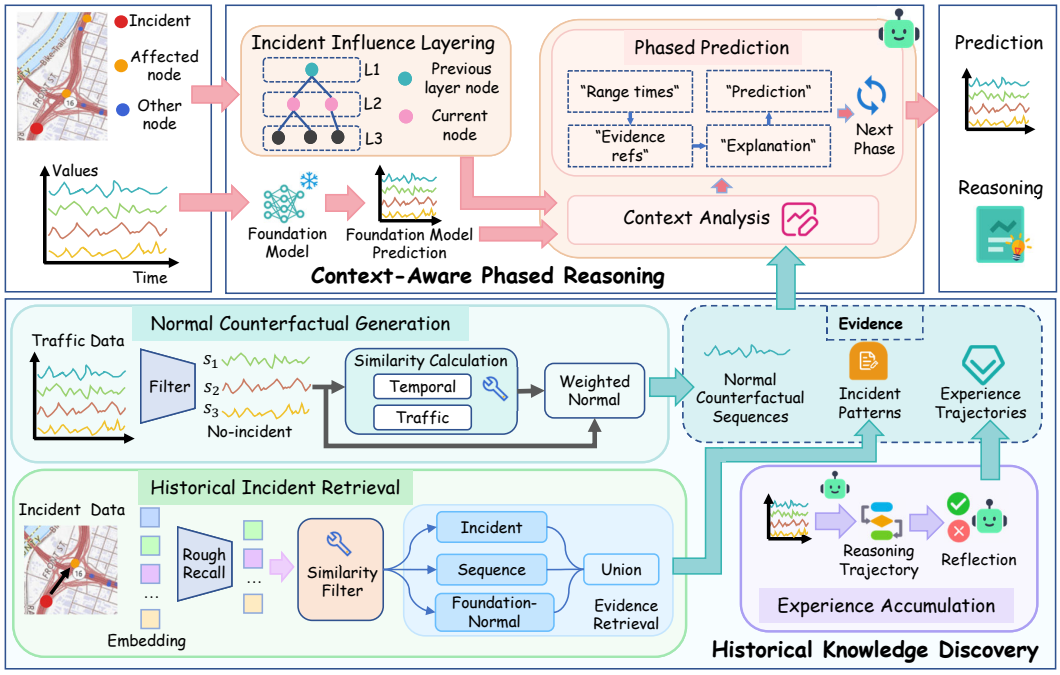


Figure 2: Framework overview.

3 PRELIMINARIES

Definition 1 (Traffic Data). The road network is represented as a weighted directed graph $\mathcal{G} = (\mathcal{V}, E, \mathbf{A})$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ denotes the set of N road sensors, E is the set of edges representing road connections, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the road adjacency matrix. Traffic data consists of multiple correlated time series collected from spatially distributed nodes where road sensors are deployed. At each time step t, the traffic observation is represented as $X^t \in \mathbb{R}^{N \times C}$, where C denotes the number of features per node (e.g., traffic flow, speed, temperature, etc.).

Definition 2 (Incident). An incident occurring at time step t_q is represented as $q = (t_q, \ell_q, r_q, d_q, c_q)$, where ℓ_q denotes latitude and longitude, r_q and d_q represent the road and direction respectively, and c_q denotes the incident type and description.

Definition 3 (Problem Statement). We define the local impact of an incident q as the incident-induced deviation from normal traffic dynamics over a subset of spatially related nodes $V_q \subseteq \mathcal{V}$, whose magnitude and temporal evolution may vary across nodes. Given the occurring incident q, the past T time steps of traffic node features $\mathbf{X} = \{X^{t-T+1}, \ldots, X^t\} \in \mathbb{R}^{T \times N \times C}$, and the affected node set V_q, we aim to learn a function f to estimate the traffic flow of the affected nodes over the next F time steps:

$$\hat{\mathbf{Y}}_{V_q} = f(q, V_q, \mathbf{X}, \mathbf{B}^q_{V_q}), \tag{1}$$

where $\hat{\mathbf{Y}}_{V_q} \in \mathbb{R}^{F \times |V_q|}$ is the prediction of traffic flow for nodes V_q, $|V_q|$ is the number of nodes in V_q, and $\mathbf{B}^q_{V_q}$ is the prediction of the foundation model.

4 METHODOLOGY

4.1 OVERVIEW

An overview of the proposed phased retrieval-augmented reasoning framework (RAP) is illustrated in Figure 2. RAP consists of two main components: Context-Aware Phased Reasoning and Historical Knowledge Discovery. Context-Aware Phased Reasoning organizes affected nodes into incident-influence layers and combines each node's foundation model prediction with retrieved evidence and predictions from previous layers. An LLM first performs context analysis and then conducts phased prediction. Historical Knowledge Discovery retrieves historical incident patterns, constructs normal counterfactual sequences from incident-free windows of the same target node, and summarizes

completed prediction episodes as reusable experience trajectories. The resulting evidence is then provided to Context-Aware Phased Reasoning for phase-consistent traffic prediction.

4.2 Context-aware Phased Reasoning

4.2.1 Incident Influence Layering

The deep learning prediction model Fang et al. (2025); Han et al. (2024); Wu et al. (2019); Guo et al. (2019) can process observations from the entire road graph through a single tensorized forward pass. Providing all affected nodes' traffic histories, foundation model predictions, and spatial relationships to a single LLM call would produce an excessively large context. RAP therefore partitions the affected node set V_q into ordered incident-influence layers $\{\mathcal{L}_1, \ldots, \mathcal{L}_K\}$ according to their driving distance relationships with the incident, with an illustrative example shown in Figure 1(b). Layers are processed sequentially, while nodes within each layer are predicted independently and can be processed in parallel. For a target node in $\mathcal{L}_k$, the LLM receives its node-specific context together with the predictions of directly connected nodes in the previous layer $\mathcal{L}_{k-1}$. This organization passes propagation evidence between layers while limiting the context required for each target node.

4.2.2 Context Analysis

The information available for each target node is heterogeneous in both form and predictive role. Traffic histories $\mathbf{X}^t_{v_i}$ describe the recent traffic state, the foundation model prediction $\mathbf{B}^q_{v_i}$ provides an initial prediction, and retrieved incident patterns $\mathcal{H}^q_{v_i}$ provide observed traffic responses under related conditions. The normal counterfactual sequence $\widetilde{\mathbf{Y}}^t_{v_i}$ represents ordinary traffic dynamics, while predictions from previous layers $\mathcal{U}^q_{v_i}$ provide propagation evidence. Retrieved experience trajectories $\mathcal{E}^q_{v_i}$ provide relevant reflections to the current context. These sources differ in relevance, reliability, and temporal alignment, and may imply conflicting traffic responses. Context Analysis evaluates these properties, reconciles conflicting signals, and consolidates the evidence into a compact reasoning context $\mathcal{R}^q_{v_i}$ for subsequent phased prediction.

4.2.3 Phased Prediction

The effect of an incident does not remain constant over the prediction horizon. Depending on the incident-to-node relationship and the evolving traffic state, its impact may be absent initially, emerge after a propagation delay, persist for several steps, or gradually diminish during recovery. RAP therefore partitions the prediction horizon into evidence-supported phases and predicts each phase using the context most relevant to its traffic condition. Building on Context Analysis, RAP generates the reasoning context $\mathcal{R}^q_{v_i}$ followed by the phased prediction $\mathcal{P}^q_{v_i}$ in a single LLM call:

$$\left(\mathcal{R}^q_{v_i}, \mathcal{P}^q_{v_i}\right) = \Phi\Big(q, \mathbf{X}^t_{v_i}, \mathcal{U}^q_{v_i}, \widetilde{\mathbf{Y}}^t_{v_i}, \mathbf{B}^q_{v_i}, \mathcal{H}^q_{v_i}, \mathcal{E}^q_{v_i}\Big), \tag{2}$$

where Φ denotes the prediction LLM. The phased prediction and the complete prediction are defined as follows:

$$\mathcal{P}^q_{v_i} = \left\{\left(s_j, e_j, \mathcal{S}_j, \widehat{\mathbf{y}}^{q,j}_{v_i}\right)\right\}_{j=1}^{J}, \qquad \widehat{\mathbf{Y}}^q_{v_i} = \widehat{\mathbf{y}}^{q,1}_{v_i} \| \cdots \| \widehat{\mathbf{y}}^{q,J}_{v_i}, \tag{3}$$

where J is the number of phases, s_j and e_j are the start and end prediction steps of phase j, $\mathcal{S}_j$ contains references to the evidence supporting that phase, and $\widehat{\mathbf{y}}^{q,j}_{v_i} \in \mathbb{R}^{e_j - s_j + 1}$ is its predicted flow sequence. The operator $\|$ denotes temporal concatenation. The phases are contiguous and cover the complete prediction horizon: $s_1 = 1$, $e_J = F$, and $s_{j+1} = e_j + 1$ for $j < J$, with $F = 12$ in our experiments.

4.3 Historical Knowledge Discovery

4.3.1 Normal Counterfactual Generation

To separate incident-induced traffic changes from ordinary traffic dynamics, RAP constructs a normal counterfactual sequence for each target node from historical traffic data under incident-free conditions. For query incident q and target node v_i, it retrieves historical windows from the same

node and filters out windows that temporally overlap with nearby incident events. For each remaining candidate window r, RAP computes a combined similarity score as follows:

$$\rho_r = \lambda \rho_r^{\text{traffic}} + (1-\lambda)\rho_r^{\text{time}}, \tag{4}$$

where ρ_r^{traffic} measures traffic-state similarity and ρ_r^{time} measures temporal similarity. Their detailed definitions are provided in Appendix A.2. RAP selects up to K_{N} candidates satisfying $\rho_r \geq \tau_{\text{N}}$. Let $\mathcal{W}_{v_i}^q$ denote the selected window set. The normal counterfactual sequence is defined as follows:

$$\widetilde{\mathbf{Y}}_{v_i}^t = \sum_{r \in \mathcal{W}_{v_i}^q} \frac{\rho_r \, \mathbf{Y}_{v_i}^r}{\sum_{r' \in \mathcal{W}_{v_i}^q} \rho_{r'}}, \tag{5}$$

where $\mathbf{Y}_{v_i}^r$ denotes the observed future flow sequence at target node v_i in historical window r. The resulting normal counterfactual sequence provides a reference for ordinary traffic dynamics.

4.3.2 Historical Incident Retrieval

Incidents are sparse and irregular events whose effects may not be adequately captured by recurring traffic patterns. RAP therefore retrieves historical incident cases to provide empirical evidence of incident-response dynamics under related conditions. To efficiently identify relevant cases from the historical corpus, RAP adopts a two-stage retrieval strategy consisting of Rough Recall and Evidence Retrieval. For a query incident q and target node v_i, the Rough Recall stage uses two complementary retrieval views to identify a broad set of potentially relevant cases. Each view retrieves up to K_{rec} cases, and their union forms the candidate pool:

$$\mathcal{C}_{v_i}^q = \text{Top}_{K_{\text{rec}}}\left(\sigma^{\text{meta}}\right) \cup \text{Top}_{K_{\text{rec}}}\left(\sigma^{\text{dyn}}\right), \tag{6}$$

where σ^{meta} measures incident-metadata similarity and σ^{dyn} measures the similarity between the query and candidate traffic histories at target node v_i. Their detailed definitions are provided in Appendix A.3. From the candidate pool $\mathcal{C}_{v_i}^q$, RAP independently retrieves three evidence groups:

$$\mathcal{H}_{v_i}^{q,\text{inc}} = \text{Top}_{K_{\text{I}}} \sigma^{\text{inc}}(q, q_j), \mathcal{H}_{v_i}^{q,\text{seq}} = \text{Top}_{K_{\text{I}}} \sigma^{\text{seq}}\left(\mathbf{X}_{v_i}^t, \mathbf{X}_{v_i}^{t_j}\right), \mathcal{H}_{v_i}^{q,\text{state}} = \text{Top}_{K_{\text{I}}} \sigma^{\text{state}}\left(\mathbf{B}_{v_i}^q, \widetilde{\mathbf{Y}}_{v_i}^t, \mathbf{B}_{v_i}^{q_j}, \widetilde{\mathbf{Y}}_{v_i}^{t_j}\right). \tag{7}$$

Here, $\mathcal{H}{v_i}^{q,\text{inc}}$, $\mathcal{H}{v_i}^{q,\text{seq}}$, and $\mathcal{H}_{v_i}^{q,\text{state}}$ contain cases with similar incident attributes, pre-incident traffic sequences, and foundation–normal states (i.e., the foundation model prediction, the normal counterfactual sequence, and their signed difference), respectively. Their union forms the Incident Patterns for target node v_i:

$$\mathcal{H}_{v_i}^q = \mathcal{H}_{v_i}^{q,\text{inc}} \cup \mathcal{H}_{v_i}^{q,\text{seq}} \cup \mathcal{H}_{v_i}^{q,\text{state}}. \tag{8}$$

Duplicate cases are retained only once in $\mathcal{H}_{v_i}^q$ while preserving their group memberships. Thus, Historical Incident Retrieval produces the incident patterns $\mathcal{H}_{v_i}^q$, where each case contains its traffic history, foundation model prediction, normal counterfactual sequence, observed future flow, and incident context. Detailed scoring and selection rules are provided in Appendix A.3.

4.3.3 Experience Accumulation

In addition to historical incident cases, RAP constructs an experience trajectory from each completed prediction episode. Each experience trajectory e is an LLM-generated record derived from an episode. Once the corresponding future observations become available, RAP evaluates the LLM prediction against the foundation model prediction as follows:

$$\Delta_e = \text{MAE}\left(\widehat{\mathbf{Y}}_{v_e}^{q_e}, \mathbf{Y}_{v_e}^{q_e}\right) - \text{MAE}\left(\mathbf{B}_{v_e}^{q_e}, \mathbf{Y}_{v_e}^{q_e}\right), \tag{9}$$

where MAE denotes the mean absolute error defined in Appendix B.1, $\widehat{\mathbf{Y}}_{v_e}^{q_e}$, $\mathbf{B}_{v_e}^{q_e}$, and $\mathbf{Y}_{v_e}^{q_e}$ denote the LLM prediction, foundation model prediction, and ground truth, respectively. A negative Δ_e indicates an improvement over the foundation model, whereas a positive value indicates degradation. Given a query incident q and target node v_i, RAP retains an experience trajectory when the foundation model error is high or the LLM prediction performs substantially worse:

$$\mathcal{M}_{v_i}^q = \{e | \, \text{MAE}\left(\mathbf{B}_{v_e}^{q_e}, \mathbf{Y}_{v_e}^{q_e}\right) \geq \theta_{\text{FM}} \ \vee \ \Delta_e \geq \theta_{\text{Fail}}\}, \tag{10}$$

where θ_{FM} and θ_{Fail} are the foundation model error and LLM error increase over the foundation model thresholds, respectively. Then, RAP compares each retained experience trajectory with the current query across the feature groups in $\mathcal{K}$, where $\mathcal{K}$ contains five feature groups: traffic history, foundation–normal state, signed foundation–normal gap, foundation model reliability, and physical relationships. For each feature group $k \in \mathcal{K}$, RAP computes the similarity and ranking score as follows.

$$\sigma_{v_i}^{q,k}(e) = \exp\left(-\frac{\left\|\mathbf{x}_{v_i}^{q,k} - \mathbf{x}_{v_e}^{q_e,k}\right\|_2}{\sqrt{d_k}}\right), \zeta_{v_i}^{q}(e) = (w_{\mathrm{static}} + w_{\mathrm{rec}}\gamma_e)\sum_{k\in\mathcal{K}} w_k\, \sigma_{v_i}^{q,k}(e). \tag{11}$$

where $\mathbf{x}_{v_i}^{q,k}$ and $\mathbf{x}_{v_e}^{q_e,k}$ are feature vectors for the query and historical experience trajectory, w_k is the weight of feature group k with $\sum_{k\in\mathcal{K}} w_k = 1$, and d_k is its feature dimension. The recency factor is $\gamma_e = 2^{-(t-t_e)/T_{\mathrm{half}}}$, where t and t_e are the prediction times of the query and trajectory e, respectively, and T_{half} is the recency half-life. The weights are nonnegative and satisfy $w_{\mathrm{static}} + w_{\mathrm{rec}} = 1$. The final experience trajectories $\mathcal{E}_{v_i}^{q}$ are the top-K_{E} trajectories ranked by $\zeta_{v_i}^{q}(e)$.

5 Experiment

5.1 Experimental Setup

All experiments are conducted on a server equipped with NVIDIA RTX 3080 GPUs, running CUDA version 12.2. All models are implemented using PyTorch. In the experiments, the proposed RAP uses AGCRN BAI et al. (2020) as its foundation model and GPT-5.6-Luna as its reasoning LLM. More details can be found in Appendix B.2.

Datasets. We conduct experiments on the large-scale TraffiDent Gou et al. (2026) benchmark dataset, which contains comprehensive traffic flow and incident records from California. All data is aggregated into 5-minute time windows. We constructed three distinct sub-datasets for our evaluation, corresponding to three different regions in California: San Francisco, Fresno, and Sacramento. More details can be found in Appendix B.4.

Baselines. We compare RAP with eleven baselines, including RNN-based, GNN-based, Attention-based, and Incident-guided methods. The RNN-based methods include AGCRN BAI et al. (2020) and DCRNN Li et al. (2018). The GNN-based methods include GWNET Wu et al. (2019), STGODE Fang et al. (2021), DGCRN Li et al. (2023), STGCN Yu et al. (2018), BigST Han et al. (2024) and DSTAGNN Lan et al. (2022). The Attention-based methods include PatchSTG Fang et al. (2025) and ASTGCN Guo et al. (2019). The Incident-guided baseline is IGSTGNN Fan et al. (2026). Detailed descriptions are provided in Appendix B.3.

Evaluation Metrics. We adopt three widely used numerical metrics to assess the quality of predicted traffic time series: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The corresponding formulas are provided in Appendix B.1.

5.2 Performance Comparison with Baselines

Table 1 reports the MAE, RMSE, and MAPE for traffic prediction across all methods on three large-scale datasets. Performance is evaluated at horizons 3, 6, and 12, as well as the average across all 12 prediction steps. RAP achieves the best average performance on all three datasets. Relative to the strongest baseline for each dataset and metric, RAP reduces average MAE, RMSE, and MAPE by 5.27%, 0.33%, and 4.13% on San Francisco; 3.65%, 3.45%, and 1.01% on Fresno; and 5.11%, 1.92%, and 3.09% on Sacramento, respectively. RAP also achieves the lowest MAE at every reported horizon on all three datasets. These results suggest that learning spatio-temporal dependencies alone leaves room for improvement under non-recurrent incidents. The incident-guided IGSTGNN incorporates incident attributes and road-network context, its average errors also remain higher than those of RAP across all three datasets. RAP further improves upon its AGCRN foundation model across all three datasets and evaluation metrics. This improvement supports augmenting learned traffic patterns with evidence relevant to the current incident. Figure 3 compares prediction errors by incident type on San Francisco. RAP reduces errors across diverse categories, including hazards, injury-related incidents, and fires, indicating that its benefits extend across different incident types.

Table 1: Traffic prediction performance comparison of the proposed RAP and baselines. Lower values indicate better predictions. **Red**: the best, Blue: the 2nd best.

Datasets	Methods	Horizon 3			Horizon 6			Horizon 12			Average		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
San Francisco	DCRNN	24.98	41.89	20.40	24.72	40.90	13.58	26.59	42.81	18.03	24.81	41.33	14.84
	GWNET	26.02	44.46	19.53	26.79	45.08	14.57	29.63	46.62	20.36	26.92	44.52	15.76
	STGODE	24.60	42.98	19.93	25.63	43.38	14.90	30.88	50.07	19.68	26.51	44.62	15.77
	DGCRN	32.09	49.99	30.82	35.92	54.42	31.47	40.84	58.01	33.93	35.56	53.10	29.30
	STGCN	24.69	41.34	20.31	24.68	41.36	14.03	27.40	43.42	18.92	25.13	41.48	15.33
	DSTAGNN	27.15	46.30	20.67	30.14	50.53	16.97	38.85	60.17	28.00	31.02	51.07	18.79
	ASTGCN	24.66	42.89	20.72	25.42	42.11	14.84	28.99	45.90	19.49	25.94	43.05	16.80
	BIGST	23.97	41.04	18.89	24.61	41.61	13.99	25.73	41.54	17.98	24.35	41.01	14.91
	IGSTGNN	24.58	41.56	18.52	24.23	40.79	**12.58**	25.44	41.10	16.78	24.63	40.92	14.03
	PatchSTG	23.81	40.88	18.79	24.20	41.45	13.74	24.96	41.01	18.24	23.80	40.75	14.72
	AGCRN	22.81	**39.70**	18.24	23.27	**40.11**	13.20	23.78	40.50	17.07	22.77	39.69	14.03
	RAP (Ours)	**22.05**	40.22	**18.04**	**21.93**	40.22	12.60	**22.30**	**40.30**	**16.53**	**21.57**	**39.56**	**13.45**
Fresno	DCRNN	17.99	32.11	19.42	18.62	30.80	18.07	21.98	32.97	20.16	19.32	31.32	21.72
	GWNET	19.10	32.58	19.61	19.99	32.59	18.92	23.23	35.66	20.94	20.60	33.20	22.47
	STGODE	18.14	31.52	19.00	18.53	30.05	17.52	21.85	33.07	20.78	19.20	31.06	20.77
	DGCRN	21.12	36.20	22.91	20.88	34.83	21.86	22.70	35.61	25.04	21.17	34.72	25.81
	STGCN	37.64	58.20	32.04	40.14	63.09	31.30	46.51	70.46	35.22	41.00	63.66	35.05
	DSTAGNN	19.96	33.28	20.32	20.79	33.14	20.61	26.04	39.17	24.79	21.66	34.27	23.37
	ASTGCN	18.74	31.98	20.96	18.85	30.53	20.67	21.62	32.30	25.92	19.51	31.23	24.54
	BIGST	17.77	30.80	18.37	18.14	29.77	17.93	19.52	29.98	17.72	18.35	29.80	20.88
	IGSTGNN	18.27	31.54	18.44	18.08	29.77	18.18	18.28	28.42	17.40	18.29	29.76	21.01
	PatchSTG	18.00	32.05	19.27	17.79	29.97	17.71	18.00	27.74	17.46	17.98	29.85	21.50
	AGCRN	17.65	31.48	18.99	17.46	29.21	17.65	17.25	26.55	17.59	17.53	28.99	21.31
	RAP (Ours)	**17.27**	**30.50**	**18.06**	**16.73**	**28.14**	**17.27**	**16.63**	**25.63**	**16.39**	**16.89**	**27.99**	**20.56**
Sacramento	DCRNN	17.08	31.75	28.69	17.86	31.61	25.28	18.93	30.32	24.94	17.44	30.09	25.71
	GWNET	17.88	32.51	28.31	19.25	33.36	25.70	20.53	33.32	26.21	18.73	32.06	26.12
	STGODE	17.33	31.95	28.62	18.29	32.06	25.47	20.34	32.75	25.80	18.16	31.18	26.15
	DGCRN	20.64	37.06	30.61	22.33	38.47	28.13	26.60	43.32	30.74	22.31	38.00	29.01
	STGCN	17.04	31.55	28.59	17.86	31.48	25.13	18.89	30.22	24.83	17.50	30.19	25.73
	DSTAGNN	18.18	32.64	28.84	20.17	34.63	26.61	24.42	40.38	30.82	20.32	34.78	28.27
	ASTGCN	17.20	31.66	27.70	18.31	32.28	24.75	19.80	31.98	24.91	17.99	30.92	25.37
	BIGST	16.71	31.31	27.20	17.21	31.01	23.66	17.12	27.77	22.03	16.64	29.15	23.95
	IGSTGNN	18.23	32.29	28.50	18.85	32.72	25.76	18.82	30.53	24.40	18.20	30.92	25.79
	PatchSTG	16.62	**31.10**	27.63	17.21	30.96	23.78	17.25	27.97	22.10	16.70	29.15	24.17
	AGCRN	17.47	31.93	28.78	17.65	31.04	24.77	17.76	28.76	23.48	17.23	29.73	25.31
	RAP (Ours)	**16.27**	31.23	**26.83**	**16.44**	**30.32**	**23.07**	**15.78**	**26.98**	**21.03**	**15.79**	**28.59**	**23.21**

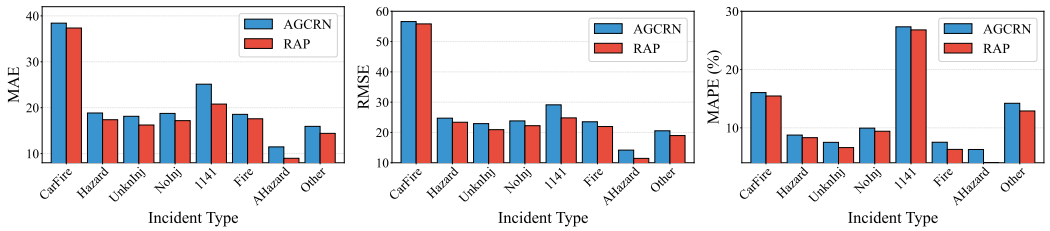


Figure 3: Performance comparison across incident types on the San Francisco dataset.

5.3 Ablation Study

To evaluate the effectiveness of RAP's components and dependence, we conduct ablation studies on San Francisco. Table 2 evaluates RAP after removing Normal Counterfactual Generation (*w/o NC*), Historical Incident Retrieval (*w/o HI*), Experience Accumulation (*w/o EA*), Phased Prediction (*w/o PP*), or Foundation Model Prediction (*w/o FP*). RAP achieves the best MAE and RMSE among these variants. Removing Foundation Model Prediction causes the largest degradation, while removing Normal Counterfactual Generation causes the next largest degradation, supporting their roles in grounding the LLM with a reference prediction and ordinary traffic evolution.

Figure 4 compares RAP with different reasoning LLMs. GPT-5.6-Luna achieves the lowest average MAE of 21.57, followed by GLM-5.3-Flash at 22.23 and Minimax-M3 at 22.37. All three outperform the AGCRN reference of 22.77. DeepSeek-V4.1-Flash and Doubao-Seed-2.0-Mini obtain higher MAEs of 22.88

Table 2: Ablation study of RAP and its variants: average results on the San Francisco dataset. **Red**: the best, Blue: the 2nd best.

Method	MAE	RMSE	MAPE (%)
w/o NC	24.52	42.42	14.23
w/o HI	22.42	39.73	13.68
w/o EA	21.82	39.77	**13.30**
w/o PP	21.93	39.68	13.40
w/o FP	25.31	43.95	14.32
RAP	**21.57**	**39.56**	13.45

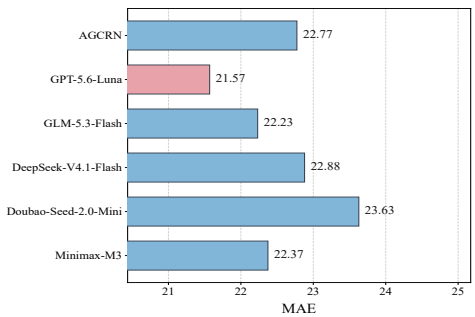


Figure 4: Average MAE results comparison of RAP with different LLMs on San Francisco.

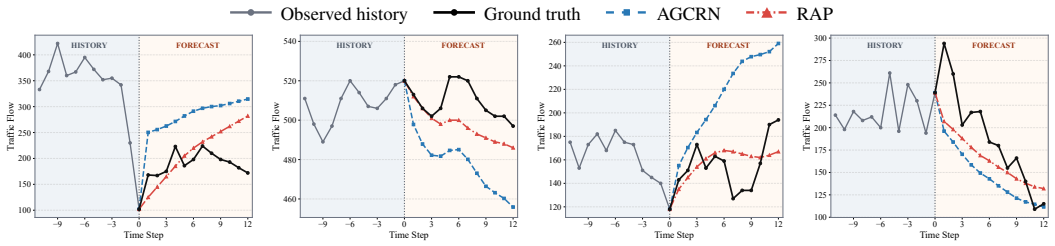
Table 3: Performance of RAP with different foundation models.

Methods	Horizon 3			Horizon 6			Horizon 12			Average		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
AGCRN	22.81	39.70	18.24	23.27	40.11	13.20	23.78	40.50	17.07	22.77	39.69	14.03
AGCRN + RAP	22.05	40.22	18.04	21.93	40.22	12.60	22.30	40.30	16.53	21.57	39.56	13.45
Improvement	↑ 3.33%	↓ 1.31%	↑ 1.10%	↑ 5.76%	↓ 0.27%	↑ 4.55%	↑ 6.22%	↑ 0.49%	↑ 3.16%	↑ 5.27%	↑ 0.33%	↑ 4.13%
ASTGCN	24.66	42.89	20.72	25.42	42.11	14.84	28.99	45.90	19.49	25.94	43.05	16.80
ASTGCN + RAP	22.88	41.60	18.77	22.97	41.18	13.21	24.65	42.86	17.49	23.17	41.27	14.59
Improvement	↑ 7.22%	↑ 3.01%	↑ 9.41%	↑ 9.64%	↑ 2.21%	↑ 10.98%	↑ 14.97%	↑ 6.62%	↑ 10.26%	↑ 10.68%	↑ 4.13%	↑ 13.15%
DGCRN	32.09	49.99	30.82	35.92	54.42	31.47	40.84	58.01	33.93	35.56	53.10	29.30
DGCRN + RAP	23.54	42.77	18.40	23.97	43.37	13.14	25.35	43.57	18.70	23.82	42.39	14.14
Improvement	↑ 26.64%	↑ 14.44%	↑ 40.30%	↑ 33.27%	↑ 20.31%	↑ 58.25%	↑ 37.93%	↑ 24.89%	↑ 44.89%	↑ 33.01%	↑ 20.17%	↑ 51.74%
IGSTGNN	24.58	41.56	18.52	24.23	40.79	12.58	25.44	41.10	16.78	24.63	40.92	14.03
IGSTGNN + RAP	22.71	41.37	18.06	22.36	41.07	12.42	22.32	39.86	16.17	22.12	40.11	13.28
Improvement	↑ 7.61%	↑ 0.46%	↑ 2.48%	↑ 7.72%	↓ 0.69%	↑ 1.27%	↑ 12.26%	↑ 3.02%	↑ 3.64%	↑ 10.19%	↑ 1.98%	↑ 5.35%
PatchSTG	23.81	40.88	18.79	24.20	41.45	13.74	24.96	41.01	18.24	23.80	40.75	14.72
PatchSTG + RAP	22.72	40.94	18.07	22.69	40.85	12.51	22.81	39.91	16.93	22.25	40.01	13.63
Improvement	↑ 4.58%	↓ 0.15%	↑ 3.83%	↑ 6.24%	↑ 1.45%	↑ 8.95%	↑ 8.61%	↑ 2.68%	↑ 7.18%	↑ 6.51%	↑ 1.82%	↑ 7.40%

and 23.63, respectively. These results show that the choice of reasoning LLM affects whether RAP can improve upon the foundation model, highlighting the importance of accurately interpreting retrieved evidence and translating it into numerical predictions.

5.4 Generalization across Foundation Models

Table 3 evaluates RAP with five foundation models on San Francisco under the same protocol. RAP consistently improves MAE and MAPE across backbones, with particularly large gains for DGCRN, despite minor degradation in a few short-horizon RMSE results. These results demonstrate RAP's effectiveness across different architectures.



(a) Hazard (b) UnknInj (c) NoInj (d) 1141

Figure 5: Case studies comparing ground truth, the AGCRN prediction, and the RAP prediction across four incident types.

5.5 Case Study

Figure 5 compares RAP with the AGCRN Base Prediction in four representative San Francisco incidents. In the Hazard case, RAP moderates the sustained increase predicted by AGCRN and more closely follows the intermediate ground-truth fluctuations. In the UnknInj case, AGCRN predicts an overly sharp decline, whereas RAP maintains a higher traffic level that remains closer to the observed trajectory over most prediction steps. For NoInj, RAP avoids the substantial upward bias of AGCRN and preserves the observed traffic range. In the 1141 case, RAP better represents the decreasing trend.

6 Conclusion

In this study, we introduce RAP, a phased retrieval-augmented reasoning framework for incident-aware traffic prediction. RAP combines a spatio-temporal foundation model with LLM-based reasoning to capture the evolving effects of traffic incidents. Historical Knowledge Discovery provides three complementary forms of evidence: incident patterns, normal counterfactual sequences, and experience trajectories. Meanwhile, Context-Aware Phased Reasoning evaluates and consolidates this evidence, then partitions the prediction horizon into impact phases to generate phase-consistent predictions. Extensive experiments on three real-world datasets demonstrate the superior performance of RAP in incident-affected scenarios.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, LLMs are used only as the reasoning and reflection modules of RAP. Note that we do not use LLMs to train the foundation model, generate benchmark data. In addition, LLMs are used for polishing the manuscript.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *ACL*, pp. 16776–16809, 2025.

Shams Forruque Ahmed, Sweety Angela Kuldeep, Sabiha Jannat Rafa, Javeria Fazal, Mahfara Hoque, Gang Liu, and Amir H Gandomi. Enhancement of traffic forecasting through graph neural network-based information fusion techniques. *Information Fusion*, 110:102466, 2024.

Ahmad Ali, Inam Ullah, Shabir Ahmad, Zongze Wu, Jianqiang Li, and Xiaoshan Bai. An attention-driven spatio-temporal deep hybrid neural networks for traffic flow prediction in transportation systems. *TITS*, 26(9):14154–14168, 2025.

Ghadah Almousa and Yugyung Lee. Traffic forecasting using spatio-temporal dynamics and attention with graph attention pdes. *Information Sciences*, 711:122108, 2025.

Ohoud Alzamzami, Zainab Alsaggaf, Reema AlMalki, Rawan Alghamdi, Amal Babour, and Lama Al Khuzayem. Passable: an intelligent traffic light system with integrated incident detection and vehicle alerting. *Sensors*, 25(18):5760, 2025.

LEI BAI, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *NeurIPS*, volume 33, pp. 17804–17815, 2020.

Jinwen Chen, Hao Miao, Chenxi Liu, Yan Zhao, and Kai Zheng. Visionst: Coordinating cross-modal traffic prediction with interactive geo-image encoding. In *WWW*, pp. 7307–7317, 2026.

Razvan-Gabriel Cirstea, Bin Yang, Chenjuan Guo, Tung Kieu, and Shirui Pan. Towards spatio-temporal aware traffic time series forecasting. In *ICDE*, pp. 2900–2913, 2022.

Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. Eta prediction with graph neural networks in google maps. In *CIKM*, pp. 3767–3776, 2021.

Hassan Eesaar, Afaq Ahmed, Muhammad Farhan, Kil To Chong, Deok Jin Lee, and Hilal Tayara. Multi-modal autonomous drone system for real-time highway incident detection and analysis. *IEEE Access*, 2025.

Lixiang Fan, Bohao Li, Tao Zou, Junchen Ye, and Bowen Du. Incident-guided spatiotemporal traffic forecasting. In *SIGKDD*, pp. 243–254, 2026.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *SIGKDD*, pp. 307–317, 2025.

Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. Spatial-temporal graph ode networks for traffic flow forecasting. In *SIGKDD*, pp. 364–373, 2021.

Ilgin Gokasar, Alperen Timurogullari, Sarp Semih Özkan, Muhammet Deveci, and Zhihan Lv. MSND: Modified standard normal deviate incident detection algorithm for connected autonomous and human-driven vehicles in mixed traffic. *TITS*, 24(11):13151–13160, 2022.

Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Bingyu Zhao, Chen Zhang, Di Wang, Xiangliang Zhang, et al. TraffiDent: A dataset for understanding the interplay between traffic dynamics and incidents. *NeurIPS*, 38, 2026.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, volume 33, pp. 922–929, 2019.

Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *PVLDB*, 17(5):1081–1090, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *ICLR*, volume 2024, pp. 23857–23880, 2024.

Georgios Kougioumtzidis, Vladimir K Poulkov, Pavlos I Lazaridis, and Zaharias D Zaharis. Mobile network traffic prediction using temporal fusion transformer. *TAI*, 6(10):2685–2699, 2025.

Mourad Lablack and Yanming Shen. Spatio-temporal graph mixformer for traffic forecasting. *Expert systems with applications*, 228:120281, 2023.

Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, and Pyang Li. DSTAGNN: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *ICML*, pp. 11906–11917, 2022.

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N Ioannidis, Huzefa Rangwala, and Christos Faloutsos. Hybgrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases. In *ACL*, pp. 879–893, 2025.

Fuxian Li, Jie Feng, Huan Yan, Guangyin Jin, Fan Yang, Funing Sun, Depeng Jin, and Yong Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *TKDE*, 17(1):1–21, 2023.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

Kenny Ye Liang, Zhongyi Pei, Huan Zhang, Yuhui Liu, Shaoxu Song, and Jianmin Wang. Retrieval-augmented generation with covariate time series. In *SIGKDD*, pp. 7624–7635, 2026.

Shichao Ma, Zhengyang Zhou, Qihe Huang, Binwu Wang, and Yang Wang. Quizsf: A retrieval-augmented framework for zero-shot time series forecasting. In *www*, pp. 1947–1958, 2026.

Boris Medina-Salgado, Eddy Sánchez-DelaCruz, Pilar Pozos-Parra, and Javier E Sierra. Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, 35:100739, 2022.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and benchmarking of llm agents: A survey. In *SIGKDD*, pp. 6129–6139, 2025.

Bokai Pan, Mingyue Cheng, Zhiding Liu, Shuo Yu, Xiaoyu Tao, Yuchong Wu, Qi Liu, Defu Lian, and Enhong Chen. Castflow: Learning role-specialized agentic workflows for time series forecasting. *arXiv preprint arXiv:2604.27840*, 2026.

Jingru Sun, Ziyu Qiu, Yichuang Sun, and Oluyomi Simpson. A self-correction transformer network for traffic flow prediction under dynamic spatio-temporal distributions. *IET ITS*, 19(1):e70044, 2025.

Xiaoyu Tao, Mingyue Cheng, Ze Guo, Shuo Yu, Yaguo Liu, Qi Liu, and Shijin Wang. Memcast: Memory-driven time series forecasting with experience-conditioned reasoning. *ICML*, 2026.

Thanh Tran, Dan He, Jiwon Kim, and Mark Hickman. MSGNN: A multi-structured graph neural network model for real-time incident prediction in large traffic networks. *Transportation research part C: emerging technologies*, 156:104354, 2023.

Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. Stone: A spatio-temporal ood learning framework kills both spatial and temporal shifts. In *SIGKDD*, pp. 2948–2959, 2024.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *IJCAI*, pp. 1907–1913, 2019.

Yaqin Ye, Yue Xiao, Yuxuan Zhou, Shengwen Li, Yuanfei Zang, and Yixuan Zhang. Dynamic multi-graph neural network for traffic flow prediction incorporating traffic accidents. *ESWA*, 234: 121101, 2023.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. A survey on evaluation of llm-based agents. In *ACL*, pp. 26690–26714, 2026.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *JICAI*, pp. 3634–3640, 2018.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *ICLR*, volume 37, pp. 11121–11128, 2023.

Weijia Zhang, Le Zhang, Jindong Han, Hao Liu, Yanjie Fu, Jingbo Zhou, Yu Mei, and Hui Xiong. Irregular traffic time series forecasting based on asynchronous spatio-temporal graph convolutional networks. In *SIGKDD*, pp. 4302–4313, 2024.

Xiyue Zhang, Chao Huang, Yong Xu, Lianghao Xia, Peng Dai, Liefeng Bo, Junbo Zhang, and Yu Zheng. Traffic flow forecasting with spatial-temporal graph diffusion network. In *AAAI*, volume 35, pp. 15008–15015, 2021.

Zijian Zhang, Xiao Han, Xiangyu Zhao, Chenjuan Guo, and Bin Yang. Learning from spatio-temporal data in the llm era: Foundations, models, and emerging trends. In *SSTD*, pp. 107–110, 2025.

Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. In *NeurIPS*, volume 38, pp. 102978–103012, 2025.

Yanliu Zheng, Juan Luo, and Mengqi Hu. Spatial-temporal incident-aware dynamic graph convolution networks for traffic flow prediction. *ESWA*, pp. 130689, 2025.

A METHODOLOGY DETAILS

A.1 ALGORITHM

Algorithm 1: RAP inference and experience update

Input: Incident q, affected nodes V_q, traffic histories, the road graph, historical incident records, and the experience trajectories.

Output: Phased predictions $\{\widehat{\mathbf{Y}}^q_{v_i}\}_{v_i \in V_q}$ and reasoning $\{\mathcal{R}^q_{v_i}\}_{v_i \in V_q}$.

Compute the Foundation Model Predictions $\{\mathbf{B}^q_{v_i}\}_{v_i \in V_q}$ for the remaining affected nodes.;

Partition V_q into ordered incident-influence layers $\{\mathcal{L}_1, \ldots, \mathcal{L}_K\}$ according to the incident–node relationships.;

for $k \leftarrow 1$ **to** K **do**

  foreach $v_i \in \mathcal{L}_k$ *in parallel* **do**

    Construct the normal counterfactual sequence $\widetilde{\mathbf{Y}}^t_{v_i}$.;

    Retrieve the candidate pool and select the incident patterns $\mathcal{H}^q_{v_i}$.;

    Retrieve the relevant experience trajectories $\mathcal{E}^q_{v_i}$.;

    Set $\mathcal{U}^q_{v_i}$ to the predictions of directly connected nodes in the previous layer.;

    $(\mathcal{R}^q_{v_i}, \mathcal{P}^q_{v_i}) \leftarrow \Phi(q, \mathbf{X}^t_{v_i}, \mathcal{U}^q_{v_i}, \widetilde{\mathbf{Y}}^t_{v_i}, \mathbf{B}^q_{v_i}, \mathcal{H}^q_{v_i}, \mathcal{E}^q_{v_i})$.;

    Concatenate the phase predictions in $\mathcal{P}^q_{v_i}$ to obtain $\widehat{\mathbf{Y}}^q_{v_i}$.;

if *episode* e *becomes available* **then**

  Evaluate and filter e, then use a reflection LLM to summarize its transferable conditions.; Store the resulting record as an Experience Trajectory.;

return $(\{\widehat{\mathbf{Y}}^q_{v_i}\}_{v_i \in V_q}, \quad \{\mathcal{R}^q_{v_i}\}_{v_i \in V_q})$.

A.2 NORMAL-COUNTERFACTUAL SIMILARITY

We detail the computation of the traffic and temporal similarities used to rank candidate normal windows. For a candidate window r at target node v_i, let $X^t_{v_i}$ and $X^r_{v_i}$ denote the query and candidate traffic values. The sequence similarity, normalized trend, and volatility are computed as follows:

$$s(r) = \exp\left(-\frac{\sqrt{\overline{\left(X^t_{v_i} - X^r_{v_i}\right)^2}}}{\overline{|X^t_{v_i}|}}\right), g(r) = \frac{\mathrm{slope}(X^x_{v_i})}{\overline{|X^t_{v_i}|}}, \nu(r) = \frac{\mathrm{std}(X^x_{v_i})}{\overline{|X^t_{v_i}|}}, x \in \{t, r\}, \tag{12}$$

where $\mathrm{slope}(\cdot)$ denotes the coefficient of a linear least-squares fit over the temporal index. Let $\mathbf{g}^x$ and $\boldsymbol{\nu}^x$ collect the corresponding descriptors over the three traffic channels. The traffic similarity is

$$\rho^{\mathrm{traffic}}_r = \mathrm{mean}(s_c(r), \exp(-\frac{\|g(t) - g(r)\|_1}{C}), \exp(-\frac{\|\nu(t) - \nu(r)\|_1}{C})), \tag{13}$$

For temporal similarity, each window is characterized by its weekday d, day period p, hour h, month m, and holiday-or-weekend status z. The cyclic similarities for hour and month are

$$s_{\mathrm{hour}}(r) = \frac{1 + \cos\left(2\pi(h_q - h_r)/24\right)}{2}, \qquad s_{\mathrm{month}}(r) = \frac{1 + \cos\left(2\pi(m_q - m_r)/12\right)}{2}. \tag{14}$$

The temporal similarity is then

$$\rho^{\mathrm{time}}_r = \mathrm{mean}(\mathbf{1}[d_q = d_r], \mathbf{1}[p_q = p_r], s_{\mathrm{hour}}(r), s_{\mathrm{month}}(r), \mathbf{1}[z_q = z_r]). \tag{15}$$

A.3 HISTORICAL INCIDENT RETRIEVAL DETAILS

This section specifies σ^{meta}, σ^{dyn}, σ^{inc}, σ^{seq}, and σ^{state} used in the main text. All traffic comparisons use the same target node v_i. The time t_j denotes the historical prediction origin associated with incident q_j, whereas t denotes the current prediction origin. Only historical cases with complete observation windows and future observations available by t are eligible; the implementation uses a 60-minute lookback cutoff for the 12-step horizon. Each recall view returns at most $K_{\mathrm{rec}} = 50$ cases. Throughout this section, Top returns case records according to the specified ranking, rather than score values.

Rough Recall. For any input vector $\mathbf{z}$, let $\mathrm{med}(\mathbf{z})$ denote its median, $\mathrm{MAD}(\mathbf{z})$ the median absolute deviation from that median, and $\mathrm{diff}(\mathbf{z})$ its adjacent first differences. Define the robust normalization Norm and dynamics encoder Dyn by

$$\begin{aligned} \mathrm{Norm}(\mathbf{z}) = \frac{\mathbf{z} - \mathrm{med}(\mathbf{z})}{\mathrm{MAD}(\mathbf{z})}, \mathrm{Dyn}(\mathbf{z}) = \mathrm{Norm}(\mathbf{z}) \| \mathrm{Norm}(\mathrm{diff}(\mathbf{z})) \| \mathrm{Wave}(\mathbf{z}), \\ \sigma^{\mathrm{meta}}(q, q_j) = (\mathbf{m}^q)^\top \mathbf{m}^{q_j}, \qquad \sigma^{\mathrm{dyn}}\big(\mathbf{X}_{v_i}^t, \mathbf{X}_{v_i}^{t_j}\big) = \mathrm{Dyn}(\mathbf{X}_{v_i}^t)^\top \mathrm{Dyn}(\mathbf{X}_{v_i}^{t_j}), \end{aligned} \tag{16}$$

where $\|$ concatenates feature blocks, and $\top$ denotes transpose. The encoder Wave applies one-dimensional wavelet scattering to $\mathrm{Norm}(\mathbf{z})$. The vectors $\mathbf{m}^q$ and $\mathbf{m}^{q_j}$ encode incident metadata: tokens from incident type, description, area, location, road, and direction are mapped to deterministic hash bins; description, area, and location also contribute character 3-grams and 4-grams. The calendar block encodes time-of-day category, weekday, weekend and holiday flags, and sine/cosine hour features, with unused positions set to zero. Thus, both recall scores are dot products of normalized embeddings, and their independently retrieved case sets form $\mathcal{C}_{v_i}^q$.

Incident Similarity. For query incident q and candidate incident q_j, let $\mathbf{m}^q$ and $\mathbf{m}^{q_j}$ denote their normalized incident-metadata vector embeddings. Let $\mathcal{F}_{\mathrm{time}}$ and $\mathcal{F}_{\mathrm{route}}$ denote the temporal and incident-to-node route attributes, respectively. The attribute, temporal, route, and overall incident similarities are

$$\begin{aligned} s_j^{\mathrm{attr}} = (\mathbf{m}^q)^\top \mathbf{m}^{q_j}, s_j^{\mathrm{time}} = \frac{1}{|\mathcal{F}_{\mathrm{time}}|} \sum_{\xi \in \mathcal{F}_{\mathrm{time}}} s_{j,\xi}^{\mathrm{time}}, \\ s_j^{\mathrm{route}} = \frac{1}{|\mathcal{F}_{\mathrm{route}}|} \sum_{\xi \in \mathcal{F}_{\mathrm{route}}} s_{j,\xi}^{\mathrm{route}}, \sigma^{\mathrm{inc}}(q, q_j) = w_{\mathrm{attr}} s_j^{\mathrm{attr}} + w_{\mathrm{time}} s_j^{\mathrm{time}} + w_{\mathrm{route}} s_j^{\mathrm{route}}, \end{aligned} \tag{17}$$

where ξ indexes an attribute in the corresponding set, and the nonnegative weights satisfy $w_{\mathrm{attr}} + w_{\mathrm{time}} + w_{\mathrm{route}} = 1$.

Here, $s_{j,\xi}^{\mathrm{time}}$ is the similarity for one temporal attribute ξ, covering holiday status, time-of-day period, and hour. Categorical attributes use agreement, while hour uses a cyclic similarity. Likewise, $s_{j,\xi}^{\mathrm{route}}$ compares one of the incident-to-node route attributes: driving distance, travel duration, or route bearing.

Sequence Similarity. For query incident q, candidate incident q_j, and target node v_i, RAP compares their pre-incident traffic histories $\mathbf{X}_{v_i}^t$ and $\mathbf{X}_{v_i}^{t_j}$. The sequence similarity reuses the traffic-state similarity $\rho_j^{\mathrm{traffic}}$ defined in normal counterfactual similarity, which compares traffic level, trend, and volatility. It additionally considers change-point and multi-scale shape evidence:

$$\sigma^{\mathrm{seq}}\left(\mathbf{X}_{v_i}^t, \mathbf{X}_{v_i}^{t_j}\right) = w_{\mathrm{traffic}} \rho_j^{\mathrm{traffic}} + w_{\mathrm{cp}} s_j^{\mathrm{cp}} + w_{\mathrm{ms}} s_j^{\mathrm{ms}}, \tag{18}$$

where the nonnegative weights satisfy $w_{\mathrm{traffic}} + w_{\mathrm{cp}} + w_{\mathrm{ms}} = 1$. The change-point representation $\mathrm{CP}(\mathbf{X})$ summarizes the dominant local transition in a traffic history, including its relative position, direction, and normalized magnitude. Its similarity is

$$s_j^{\mathrm{cp}} = \cos\big(\mathrm{CP}(\mathbf{X}_{v_i}^t), \mathrm{CP}(\mathbf{X}_{v_i}^{t_j})\big), \tag{19}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. Let $\mathcal{N}$ denote a set of temporal resolutions, and let $\mathrm{Wave}_\omega(\mathbf{X})$ denote the wavelet or scattering representation of $\mathbf{X}$ at resolution $\omega \in \mathcal{N}$. The multi-scale shape similarity is

$$s_j^{\mathrm{ms}} = \sum_{\omega \in \mathcal{N}} w_\omega^{\mathrm{ms}} \cos\big(\mathrm{Wave}_\omega(\mathbf{X}_{v_i}^t), \mathrm{Wave}_\omega(\mathbf{X}_{v_i}^{t_j})\big), \qquad \sum_{\omega \in \mathcal{N}} w_\omega^{\mathrm{ms}} = 1. \tag{20}$$

Thus, $\rho_j^{\mathrm{traffic}}$ captures the shared traffic state, s_j^{cp} captures local transition behavior, and s_j^{ms} captures traffic shapes across temporal resolutions. Only pre-incident histories are used for ranking; observed future flow is excluded from σ^{seq}.

Foundation–Normal state similarity. This group compares the Foundation Model Prediction, the Normal counterfactual sequence, and their signed difference. Let $\kappa \in \mathcal{A} = \{\mathrm{pred}, \mathrm{normal}, \mathrm{gap}\}$ denote the trajectory type (prediction, normal counterfactual, and their difference). For node v_i, the query and historical trajectories are defined as:

$$\mathbf{z}_{v_i}^{q,\kappa} = \begin{cases} \mathbf{B}_{v_i}^{q}, & \kappa = \mathrm{pred}, \\ \widetilde{\mathbf{Y}}_{v_i}^{t}, & \kappa = \mathrm{normal}, \\ \mathbf{B}_{v_i}^{q} - \widetilde{\mathbf{Y}}_{v_i}^{t_j}, & \kappa = \mathrm{gap}, \end{cases} \tag{21}$$

where $\mathbf{B}$ denotes the Foundation Model Prediction, $\widetilde{\mathbf{Y}}$ denotes the Normal counterfactual sequence, and the gap trajectory preserves their signed difference. Let $\overline{|\mathbf{X}_{v_i}^{t}|}$ denote the average absolute value of the query history. This shared scale normalizes the discrepancies between corresponding query and candidate sequences:

$$\begin{aligned} d_j^{\kappa,\mathrm{value}} &= \frac{\left\|\mathbf{z}_{v_i}^{q,\kappa} - \mathbf{z}_{v_i}^{q_j,\kappa}\right\|_2}{\overline{|\mathbf{X}_{v_i}^{t}|}\sqrt{\dim(\mathbf{z}_{v_i}^{q,\kappa})}}, \\ d_j^{\kappa,\mathrm{change}} &= \frac{\left\|\mathrm{diff}\left(\mathbf{z}_{v_i}^{q,\kappa}\right) - \mathrm{diff}\left(\mathbf{z}_{v_i}^{q_j,\kappa}\right)\right\|_2}{\overline{|\mathbf{X}_{v_i}^{t}|}\sqrt{\dim(\mathrm{diff}(\mathbf{z}_{v_i}^{q,\kappa}))}}. \end{aligned} \tag{22}$$

Here, $\|\cdot\|_2$ is the Euclidean norm, $\dim(\cdot)$ returns the vector dimension, and diff returns adjacent first differences. Let $\mathcal{D} = \mathcal{A} \times \{\mathrm{value}, \mathrm{change}\}$ index all discrepancy types. Arrange $\{d_j^{\kappa,\delta} : (\kappa,\delta) \in \mathcal{D}\}$ in nonincreasing order as $d_{j,(1)} \geq \cdots \geq d_{j,(|\mathcal{D}|)}$. The state ranking quantity is the ordered tuple

$$\sigma^{\mathrm{state}}\left(\mathbf{B}_{v_i}^{q}, \widetilde{\mathbf{Y}}_{v_i}^{t}, \mathbf{B}_{v_i}^{q_j}, \widetilde{\mathbf{Y}}_{v_i}^{t_j}\right) = \left(-d_{j,(1)}, \ldots, -d_{j,(|\mathcal{D}|)}\right). \tag{23}$$

The $\mathrm{Top}_{K_\mathrm{I}}$ operator compares these tuples in descending lexicographic order, prioritizing the smallest worst-case discrepancy and then the subsequent discrepancies. This multi-dimensional ordering prevents a strong match in one trajectory from hiding a severe mismatch in another.

B Details of the Experiment Setup

B.1 Evaluation metrics

We use three widely adopted numerical metrics to assess the quality of predicted traffic time series: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The formulas for these metrics are:

$$\mathrm{MAE} = \frac{1}{TN}\sum_{j=1}^{T}\sum_{i=1}^{N}\left|\hat{y}_i^j - y_i^j\right|, \tag{24}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{TN}\sum_{j=1}^{T}\sum_{i=1}^{N}\left(\hat{y}_i^j - y_i^j\right)^2}, \tag{25}$$

$$\mathrm{MAPE} = \frac{1}{TN}\sum_{j=1}^{T}\sum_{i=1}^{N}\left|\frac{\hat{y}_i^j - y_i^j}{y_i^j}\right| \times 100\%, \tag{26}$$

where $\hat{y}_i^j$ denotes the predicted value for the i-th node at the j-th time step, and y_i^j represents the actual value of the i-th node at the j-th time step. Here, T refers to the total number of time steps, and N represents the total number of nodes.

B.2 Implementation Details

Data preparation and evaluation. Traffic observations are sampled every five minutes. We construct sliding windows with $T = 12$ historical steps and $F = 12$ future steps, corresponding to one hour of history and one hour of prediction. The windows are ordered chronologically and divided

Table 4: Experimental parameters

Parameter	Value	Description
Normal Counterfactual Generation		
K_{N}	10	Number of selected incident-free windows.
τ_{N}	0.6	Threshold for selection.
λ	0.75	Weight of traffic-state similarity.
Historical Incident Retrieval		
K_{rec}	50	Number of recalled candidates .
K_{I}	3	Number of retrieved cases per evidence group.
Experience Accumulation		
K_{E}	3	Number of retrieved experiences.
θ_{FM}	15	Threshold for foundation model prediction MAE.
θ_{Fail}	10	Threshold for LLM error increase over the foundation model.
T_{half}	30 days	Half-life of the recency factor.
w_{static}	0.9	Static weight in the recency multiplier.
w_{rec}	0.1	Weight of the time-dependent recency factor.

into training, validation, and test sets in a 70%/15%/15% ratio. For TraffiDent, the inputs comprise flow, occupancy, and speed, and the prediction target is flow. Incident evaluation uses incidents whose prediction windows belong to the test split and whose affected nodes fall within the configured driving-distance scope. Predictions are evaluated in the original units, excluding zero-valued targets that represent missing observations.

Foundation Model training. All traffic prediction models used as Foundation Models follow the same training protocol. Each model is optimized with Adam using an L_1 loss, a learning rate of 10^{-3}, and weight decay of 10^{-4} for 80 epochs. The checkpoint with the lowest validation MAE is selected for testing. After training, the Foundation Model is kept fixed during RAP inference, while GPT-5.6-Luna performs the LLM-based reasoning.

Incident Influence layering details. We use Open Source Routing Machine (OSRM) with the driving profile to obtain incident-to-sensor route distances and travel durations, which are precomputed and cached. OSRM snaps these coordinates to the road network; route bearings are computed from the snapped coordinates. The affected node set V_q contains sensors within 1.5 km driving distance of the incident. Layers $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ correspond to distance intervals $[0, 0.5]$, $(0.5, 1.0]$, and $(1.0, 1.5]$ km, respectively. Layers are processed sequentially, while nodes within each layer can be processed in parallel. Incidents without a reachable sensor in this scope are excluded from incident evaluation.

Experimental parameters. Table 4 shows the experimental parameters.

B.3 Details of Baselines

We compare the proposed method with a comprehensive set of baselines, including RNN-based, GNN-based, Attention-based, and Incident-guided methods.

- **STGCN** Yu et al. (2018) combines graph convolution with gated temporal convolution to jointly model spatial and temporal dependencies.
- **DCRNN** Li et al. (2018) integrates diffusion graph convolution with recurrent neural networks to capture directional spatial dependencies and temporal dynamics.
- **AGCRN** BAI et al. (2020) learns node-specific graph convolution filters and adaptive node embeddings to capture heterogeneous spatial dependencies.
- **ASTGCN** Guo et al. (2019) incorporates spatial and temporal attention mechanisms into graph convolution to capture dynamic spatio-temporal dependencies.
- **GWNET** Wu et al. (2019) combines dilated temporal convolution with diffusion graph convolution and learns a self-adaptive adjacency matrix.

- **STGODE** Fang et al. (2021) employs ordinary differential equations to continuously model the evolution of spatio-temporal traffic dynamics.
- **DSTAGNN** Lan et al. (2022) introduces dynamic spatial-temporal attention to capture time-varying spatial dependencies and complex traffic interactions.
- **DGCRN** Li et al. (2023) integrates a dynamic graph with a predefined static graph to model evolving spatial dependencies.
- **PatchSTG** Fang et al. (2025) adopts patch-based representations to efficiently model spatial dependencies for large-scale traffic prediction.
- **BigST** Han et al. (2024) efficiently exploits long-range spatio-temporal dependencies for large-scale traffic prediction.
- **IGSTGNN** Fan et al. (2026) incorporates incident attributes, road-network relationships, and contextual information into spatio-temporal prediction.

Table 5: Dataset Statistics

Dataset	Nodes	Edges	Incidents	Samples	TimeSlices	Timespan
San Francisco	76	1,501	9,701	8M	105,408	2024-01-01 – 2024-12-31
Fresno	334	10,724	7,657	35M	105,408	2024-01-01 – 2024-12-31
Sacramento	836	36,358	21,936	88M	105,408	2024-01-01 – 2024-12-31

Table 6: Description of Incident Information Features.

Feature Name	Type	Example
Location	String	I80 E / The Embarcadero Uc
Description	String	1125-Traffic Hazard
Type	String	Hazard
Latitude, Longitude	Float	37.789053, -122.387523
Absolute Postmile	Float	119.3
Freeway Name	String	SR-99

Table 7: Taxonomy and Definitions of Incident Types.

Incident Type	Typical Examples	Key Definitions / Clarification
Hazard	1125-Traffic Hazard, HAZMAT-Hazardous Materials Inc	General non-collision traffic hazards, including temporary road obstructions and hazardous materials events.
NoInj	1182-Trfc Collision-No Inj, 20002-Hit and Run No Injuries	Traffic collisions or hit-and-run events with no reported injuries.
UnknInj	1183-Trfc Collision-Unkn Inj	Traffic collisions where injury status is unknown or not yet determined.
1141	1179-Trfc Collision-1141 Enrt	Traffic collisions with ambulance en route.
Fire	FIRE-Report of Fire	Reported fires other than vehicle-specific car fires, such as roadside or vegetation fires.
CarFire	CFIRE-Car Fire	Vehicle fires.
AHazard	1125A-Animal Hazard, ANIMAL-Live or Dead Animal	Animal-related roadway hazards, including live or dead animals.
Other	1013-Road/Weather Conditions, 1144-Fatality, 1166-Defective Traffic Signals, 1180-Trfc Collision-Major Inj	Miscellaneous CHP incident codes not mapped to the categories above, including weather, signals, fatalities, and injury collisions.

B.4 Details of datasets

Table 5 summarizes the dataset statistics. Table 6 presents representative incident and road-location attributes, including incident type, description, geographic coordinates, and road identifiers. Table 7 summarizes the eight incident categories and explains their reporting codes.

C More Experimental Results

C.1 Baseline Results on the Full Test Set

Table 8 reports the baseline results on the complete chronological test split, rather than on the incident-only evaluation subset used in Table 1. The test split contains all valid sliding windows in the final 15% of each regional time series.

Table 8: Baseline performance on the full chronological test set. Lower values are better.

Dataset	Method	Horizon 3			Horizon 6			Horizon 12			Average		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
San Francisco	AGCRN	14.67	26.30	8.59	16.32	29.11	9.40	18.86	33.11	10.83	16.27	29.10	9.46
	DCRNN	15.92	27.17	9.10	17.86	30.14	9.77	21.17	34.60	11.61	17.89	30.17	9.95
	GWNET	17.09	28.66	9.54	20.06	33.00	10.96	24.97	40.01	13.83	20.14	33.39	11.19
	STGODE	15.98	27.88	9.31	18.51	31.91	10.66	23.50	39.05	12.95	18.75	32.37	10.69
	DGCRN	22.94	34.85	16.31	28.24	41.86	20.25	35.17	51.16	22.78	27.98	42.09	19.46
	STGCN	16.18	27.40	9.64	18.33	30.60	10.36	22.32	35.81	12.59	18.45	30.75	10.61
	DSTAGNN	18.31	30.58	10.63	22.08	36.67	12.39	30.55	48.06	19.39	22.79	37.89	13.46
	ASTGCN	16.21	27.83	9.62	18.52	31.38	10.71	22.92	37.03	13.21	18.75	31.61	11.20
	BIGST	15.31	26.87	8.83	17.77	30.15	10.07	21.50	35.29	11.94	17.78	30.36	10.09
	IGSTGNN	16.08	27.09	8.95	17.76	29.71	9.74	20.86	34.19	11.51	17.91	29.93	9.91
	PatchSTG	15.05	26.63	8.62	17.00	29.50	9.62	19.80	33.34	11.29	16.94	29.44	9.69
Fresno	AGCRN	11.00	17.15	17.89	11.72	18.48	18.73	13.13	20.84	21.21	11.79	18.61	18.94
	DCRNN	10.83	16.71	19.05	11.62	18.13	19.97	13.39	20.96	23.16	11.76	18.37	20.38
	GWNET	12.15	19.08	19.14	13.66	21.85	20.73	16.11	26.15	23.67	13.71	22.14	20.83
	STGODE	11.62	17.99	18.47	12.94	20.35	20.05	16.19	25.80	24.17	13.25	21.06	20.46
	DGCRN	12.94	19.56	27.77	14.59	23.35	29.71	18.41	30.24	35.98	15.04	24.28	30.20
	STGCN	11.25	17.49	18.07	12.33	19.48	19.26	14.86	23.68	22.63	12.55	19.92	19.61
	DSTAGNN	13.07	20.62	20.03	14.95	23.86	23.53	19.49	31.12	29.52	15.36	24.72	23.78
	ASTGCN	11.62	18.07	19.06	12.68	20.12	20.43	15.15	25.06	24.08	12.87	20.78	20.79
	BIGST	11.01	17.12	17.84	12.10	19.13	19.37	14.20	22.67	21.87	12.17	19.32	19.40
	IGSTGNN	11.68	18.28	18.09	12.18	19.03	19.46	13.75	21.69	20.92	12.30	19.32	19.31
	PatchSTG	10.88	16.93	17.81	11.64	18.38	18.56	13.14	20.95	20.31	11.72	18.53	18.72
Sacramento	AGCRN	11.06	18.59	24.97	11.85	19.91	26.33	13.54	22.74	29.00	11.91	20.07	26.42
	DCRNN	10.56	17.82	23.77	11.55	19.54	25.16	13.54	22.80	28.39	11.66	19.79	25.53
	GWNET	11.47	19.35	24.87	13.05	22.11	27.19	15.45	26.13	31.28	13.06	22.31	27.35
	STGODE	11.16	18.85	25.06	12.54	21.20	27.69	15.46	26.04	31.57	12.73	21.68	27.58
	DGCRN	12.99	21.99	30.63	15.11	25.45	35.42	21.19	35.64	47.01	15.81	27.10	36.40
	STGCN	10.81	18.28	24.85	11.96	20.22	26.75	14.47	24.40	30.30	12.15	20.67	26.88
	DSTAGNN	12.19	20.49	25.99	14.20	23.94	29.58	18.62	31.00	38.05	14.55	24.72	31.74
	ASTGCN	11.00	18.59	24.04	12.41	20.95	26.86	14.98	25.16	30.66	12.52	21.29	26.78
	BIGST	10.41	17.74	22.67	11.22	19.13	24.10	12.73	21.49	26.36	11.26	19.20	24.15
	IGSTGNN	11.61	19.65	24.58	12.43	20.99	26.67	14.06	23.53	29.56	12.44	21.04	26.55
	PatchSTG	10.38	17.60	23.47	11.27	19.12	24.55	12.67	21.38	26.81	11.30	19.20	24.72

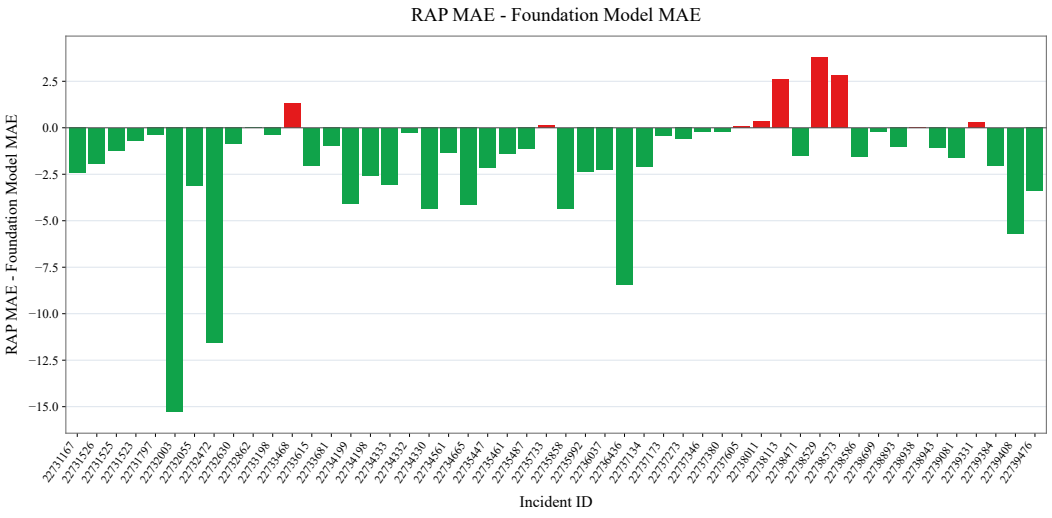


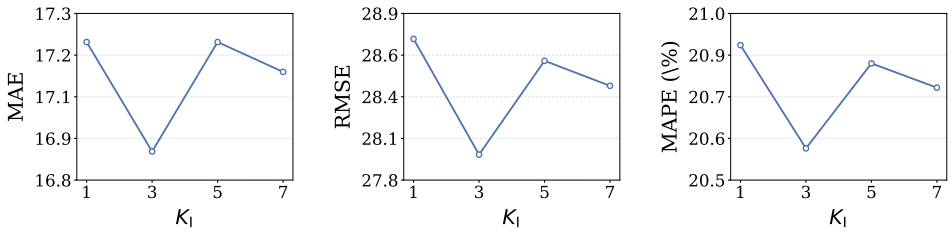
Figure 6: Incident-level MAE differences between RAP and the foundation model for the first 50 evaluated incidents. Negative values (green) indicate lower MAE with RAP, while positive values (red) indicate higher MAE.

C.2 Performance Visualization

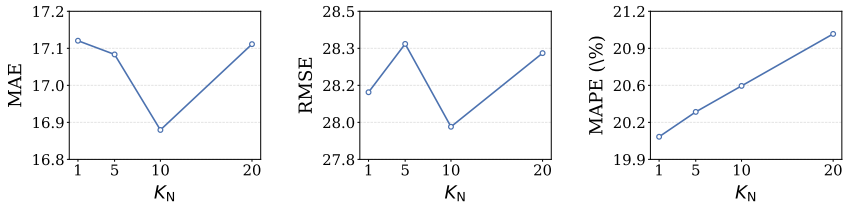
Figure 6 compares RAP with the foundation model on the first 50 evaluated incidents in the San Francisco dataset, where negative MAE differences indicate improvement. RAP reduces MAE in 41 of the 50 incidents (82%), demonstrating improvements across a broad range of incidents.

C.3 Parameter Sensitivity

Figure 7 evaluates the number of retrieved cases per evidence group K_{I} and the number of selected incident-free windows K_{N}. The default settings, $K_{\mathrm{I}} = 3$ and $K_{\mathrm{N}} = 10$, achieve the lowest MAE.



(a) The number of retrieved cases per evidence group.



(b) The number of selected incident-free windows.

Figure 7: Parameter sensitivity on the Fresno dataset.

C.4 Efficiency Analysis

Table 9 reports the runtime on San Francisco incidents. Parallel processing within each incident-influence layer reduces the mean per-incident time from 270.12 to 128.84 seconds, a 52.3% reduction. The per-node time remains 75.58 seconds because layering changes node scheduling rather than the cost of an individual prediction.

Table 9: Runtime on the San Francisco dataset.

Method	Per-Incident Time (s)	Per-Node Time (s)
Without Layering	270.12	75.58
Layering	128.84	75.58

C.5 Token Consumption

Table 10 summarizes the LLM calls and token usage across incident-influence layers. Layer 2 and Layer 3 account for 44.2% and 40.0% of the total tokens, respectively, while Layer 1 accounts for 15.7%.

Table 10: LLM token consumption by incident-influence layer.

Metric	Layer 1	Layer 2	Layer 3	**Total**
Prediction calls	382	1,012	929	**2,323**
Reflection calls	332	955	831	**2,118**
Total tokens	10.37M	29.12M	26.37M	**65.87M**

D SHOW CASE

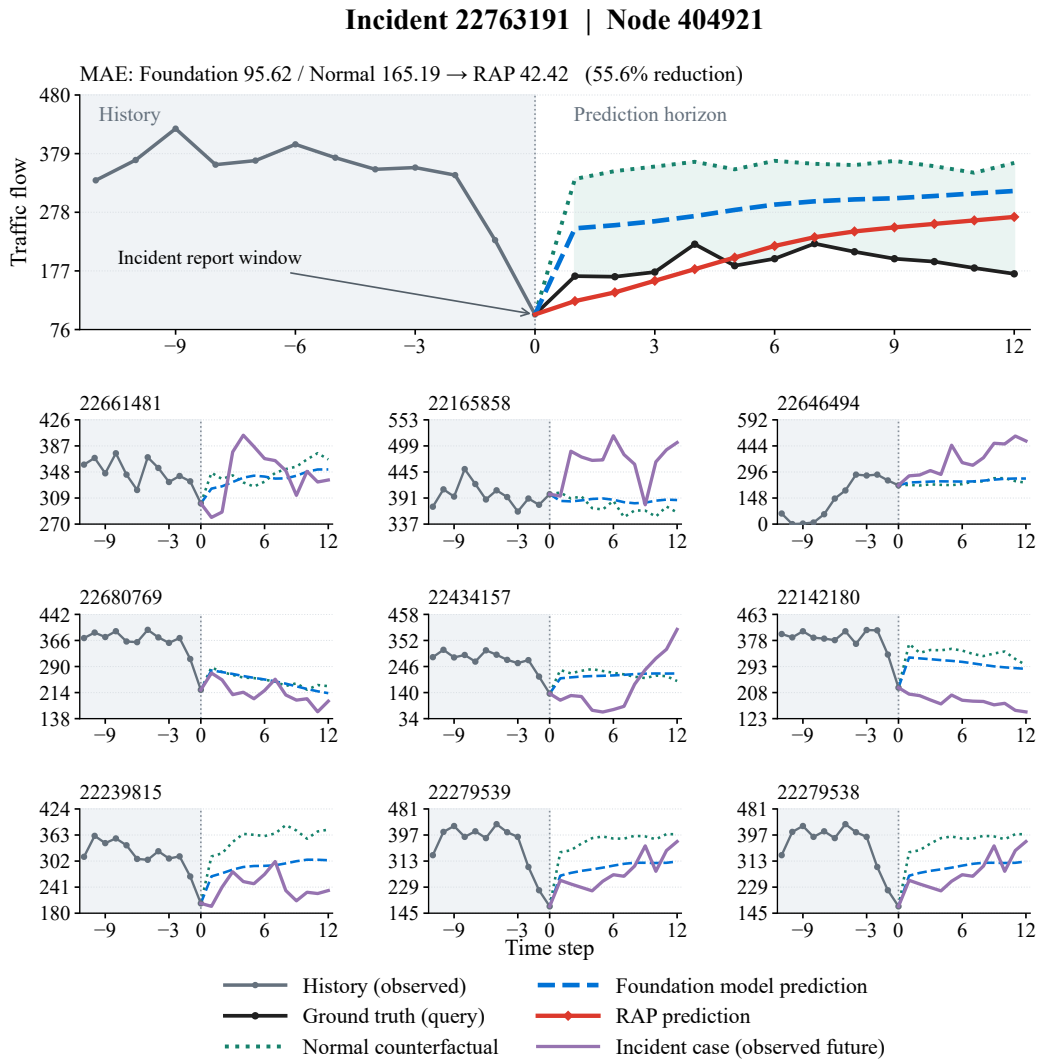


Figure 8: RAP prediction and retrieved incident cases for incident 22763191 at node 404921.

E PROMPT DETAILS

Prompt 1: Incident-Aware Phased Traffic Prediction Prompt

[System message]
You are an incident−aware traffic flow prediction agent. Use only the supplied context and do not invent unavailable evidence. Follow the task and the output requirements below.
Output requirements
− Return one JSON object with exactly ‘target_node_id‘, ‘predict_explanation‘, and ‘phase_analysis‘, in that order.
− ‘predict_explanation‘ is a brief natural−language string describing the key evidence and important uncertainty. Mention cases when they materially affect the decision; no separate analytical fields or candidate tables are required.
− ‘phase_analysis‘ is the final field and has 1−4 contiguous phases covering steps 1−12 without gaps or overlap. Every phase contains ‘start_step‘, ‘end_step‘, ‘evidence_references‘, ‘explanation‘, and ‘predicted_flow‘. These are the following four semantic parts, in this order: ‘Range times‘ (‘start_step‘ and ‘end_step‘, inclusive); ‘Evidence refs‘ (‘evidence_references‘, a JSON list identifying supplied evidence that materially supports the phase); ‘Explanation‘ (‘explanation‘, a concise account of how that evidence supports or limits the phase); and ‘Prediction‘ (‘predicted_flow‘, the phase’s flow values). Include all four parts in every phase; use an empty list only when no specific source applies for ‘evidence_references‘. Predicted values are finite, non−negative, exactly the phase length, and rounded to at most two decimals.
− Do not rename keys, add wrappers or extra top−level keys, return a second JSON object, or append another explanation after the phase predictions. The phase object in the skeleton is illustrative only: replace its placeholders with one or more evidence−supported phases covering the full horizon.
Required JSON top−level skeleton
{”target_node_id”: <target_node_id>,”predict_explanation”: ”...”, ”phase_analysis”: [{”start_step”: ”<start_step>”,”end_step”: ”<end_step>”, ”evidence_references”: [],”explanation”: ”<phase explanation>”, ”predicted_flow”: [”<one value per step in this phase>”]}]}
[User message]
Task:Predict all 12 future five−minute steps for the target node from the supplied context. Compare the evidence and competing future hypotheses, then construct the flow sequence that best fits this sample. The resulting shape can be non−uniform or non−linear when supported by the evidence.
Evidence principles
1. Current observations. The query is the current incident and target node; ‘traffic_history‘ and ‘candidate_history_flow‘ from a case are observed pre−predict states, never future observations for the query.
2. Hypotheses. ‘foundation_model_prediction‘ is a history−conditioned baseline learned without an incident indicator; it is neither a strict no−incident counterfactual nor an incident−conditioned prediction and may already contain effects visible in the history. ‘normal_counterfactual_sequence‘ is the conditional future of matched incident−free windows. Both are candidate descriptions of the future, not observations or truth. The signed Foundation−Normal gap is the Foundation Model Prediction minus the Normal counterfactual sequence at each step; ‘foundation_normal_gap.G‘ is its mean absolute separation, not its direction. The gap is derived, not independent evidence. Historical Foundation Model errors are errors on matched incident−free windows: horizon bands are step−wise error statistics, not phases, and positive bias means the Foundation Model Prediction overpredicted. Normal spread is historical variation, not calibrated confidence.
3. Historical outcomes. ‘historical_effect.observed_future_flow‘ and ‘paired_state.historical_outcome.observed_future_flow‘ are observed futures after a historical case pre−prediction state. ‘effect_residual‘ is that future minus the local no−incident reference for the case; paired outcomes also contain observed−minus−Foundation−Model and observed−minus−Normal residuals. Null observations do not establish accuracy. These residuals and before/after observations are uncertain analog evidence, never causal estimates. ‘trajectory_comparisons‘ separates signed level offsets, centered shape changes, and step changes; differences are in flow units, and gap differences (Foundation Model differences minus Normal differences) are not independent votes.
4. Historical Incident Retrieval evidence. Incident Patterns contain three evidence groups: ‘sequence_similarity‘, ‘incident_similarity‘, and ‘foundation_normal_state_similarity‘. They have different evidence roles. A case may inform only level, direction, amplitude, persistence, a turning point, or one interval; rank and similarity do not establish transferability. The normalized ‘retrieval_score‘ uses time, cached OSRM route, target history, and incident attributes. Target−history similarity uses wavelet−scattering shape, robust level, latest three steps, and

changepoints. The Foundation−Normal state similarity compares the Foundation Model Prediction, Normal counterfactual sequence, and signed gap jointly and minimizes the worst normalized mismatch before the next worst, so one strong component cannot hide a severe mismatch. Group, rank, score, counts, and programmatic direction/shape aggregates are descriptive metadata, not votes or reliability estimates.

5. Physical and memory evidence. Cached OSRM driving route, road relation, topology, target history, and time context inform whether and when an incident could reach the target. Reachability is only a possible exposure and timing signal: it does not establish the sign, magnitude, or persistence of a flow effect. An incident adjustment needs compatible target−state outcomes or an already observed target change; route closeness alone is insufficient. Previous−layer predictions are weak cross−node evidence only: node−scale values differ, so an upstream level or trajectory is not the target prediction and does not prove propagation direction or size. Experience Trajectories are conditional historical prediction evidence, not a prescription or causal estimate.
6. Rules shared by all sources. Evaluate every enabled source against the complete sample; preserve conflict, weak coverage, and competing explanations. Do not majority−vote, mechanically average cases, or treat a historical winner as decisive. When conflicting outcomes cannot be distinguished by observable pre−prediction features, keep the result unresolved rather than inventing a discriminator or selecting the most extreme case. Judge transferability from observable query−case features and its limitations. Preserve competing explanations when the evidence does not distinguish them. Infer freely: do not copy an existing trajectory, force the result between the Foundation Model Prediction and Normal, average them, or prefer either when analogs are weak. Context Analysis is the evidence consolidation step; Phased Prediction is the resulting horizon−structured prediction in the same LLM response.

Reasoning process

1. Establish the current state from the complete target history: level, overall and recent trend, volatility, turning points, and likely persistence. Treat the latest direction as a hypothesis, then test whether it is likely to persist, reverse, or develop further against the full history, the Foundation Model Prediction, the Normal counterfactual sequence, and matched−case continuations to form the current−state hypothesis. The latest one−to−three observations alone must not determine future amplitude or persistence. Use time of day, weekday, comparable windows, and ordinary traffic evolution; the incident contribution may be small or unresolved.
2. Develop competing futures from the Foundation Model Prediction, Normal counterfactual sequence, signed Foundation−Normal gap, and structural case summaries. Compare them with the current−state hypothesis, use ‘G‘ and raw trajectories to locate differences, and use historical Foundation Model error summaries to judge matched−window reliability. Check the premise and history−match quality of the Normal counterfactual. Separate ordinary demand evolution from an incident effect: the cached OSRM driving route (‘incident.osrm_driving_route.distance_km‘, ‘duration_s‘), road relation, and topology constrain possible reachability and timing, but do not determine effect sign or magnitude without compatible outcome evidence. Include target history and connected previous−layer predictions only with their transfer limits; never use straight−line distance. Report time may lag onset: do not assume a fixed delay or invent a time, and allow effects already visible in history.
3. Evaluate only retained Incident Patterns and Experience Trajectories. Inspect group scores, ranks, retrieval and OSRM components, ‘effect_shape‘, ‘effect_residual‘, direction, spread, per−step distributions, and ‘trajectory_comparisons‘; inspect raw paths when summaries hide a turning point or local mismatch. For the Sequence Similarity group, compare ‘candidate_history_flow‘ with the query ‘traffic_history‘, then its future continuation: turning points, short deviations, recovery, direction, ordering, persistence, relative magnitude, and timing. Adapt it to target level, recent dynamics, calendar, and physical evidence; road relation does not determine transferability alone. For every case, separate pre−prediction state similarity from outcome transfer: a mismatch in flow scale, timing, recent shape, or incident duration limits the transferred amplitude even when the route is close. For Incident Similarity cases, use ‘effect_residual‘ to assess a possible incident mechanism and interval transfer from incident/route context, reachability, node scale, and residuals. For Foundation−Normal State Similarity cases, compare the Foundation Model Prediction, Normal counterfactual sequence, and signed gap together before using the outcome. Judge each case independently by interval, allowing support or conflict; use memory with the same transfer limitations.
4. Synthesize plausible continuations from all enabled sources. For each hypothesis and case, identify supporting and contradicting conditions, historical performance, transferable features, and limits. Treat evidence as support, counterexample, or unresolved rather than a vote. Explain which observable query−case features transfer level, direction, persistence, magnitude, or turning time. When cases conflict, compare the strongest supporting case and counterexample, but use a

distinguishing feature only when it is observable in the supplied pre−prediction state. If no such feature separates the outcomes, explicitly preserve the interval as unresolved. A single extreme historical outcome, a close route, or a high rank does not set the prediction amplitude or persistence. Select the explanation that best accounts for each interval from convergent evidence, while allowing competing explanations to remain unresolved.

5. Construct phases and flow values. Before assigning values, state internally whether an incident effect is established, directionally supported, or unresolved for each interval; do not invent an incident sign or magnitude when the evidence only establishes reachability. Divide steps 1−12 only where supported by the future evolution; the reported incident time is not a phase boundary. Relate evidence to each phase and connect direction and magnitude to the explanation. Before values, give a concise explanation naming material cases by incident ID and citing query/case features, timing or road context, Foundation−Normal/gap differences, historical performance, transfer value, limits, counterevidence, and uncertainty. Summarize decisions without repeating the full evidence review, then generate the flow sequence.

Input context

<model−facing JSON containing the incident, traffic history, Foundation Model Prediction and Normal counterfactual trajectories when enabled, retrieved Incident Patterns, connected previous−layer predictions, and Experience Trajectories when enabled>

Prompt 2: Experience Reflection Prompt

[System message]

You are an Experience Trajectory reviewer. Review one completed incident−aware traffic prediction after its outcome is known. Write one concise paragraph stating the main agreement or error between the prediction and the observed outcome, one observable condition under which the experience may transfer, and its limitation. Foundation Model Prediction residuals are prediction errors, not measured causal incident effects. Use only the supplied evidence, record observations rather than commands, and do not repeat full trajectories or invent causes.